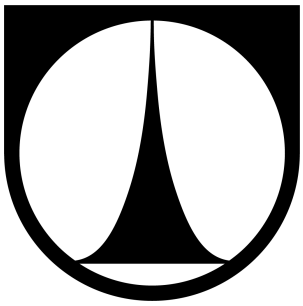
TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky, informatiky a mezioborových studií



Bakalářská práce

Liberec 2013 **Slavík Václav**

TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky, informatiky a mezioborových studií

Studijní program : B2646 – Informační technologie
Studijní obor: 1802R007 – Informační technologie

On-line systém pro modelování a predikci sportovních výsledků

On-line system for modeling and predicting sports results

Bakalářská práce

Autor:	Slavík Václav
Vedoucí práce:	doc. Petr Volf, CSc., prom. mat.

V Liberci 17. 5. 2013

Prohlášení

Byl(a) jsem seznámen(a) s tím, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu TUL.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Bakalářskou práci jsem vypracoval(a) samostatně s použitím uvedené literatury a na základě konzultací s vedoucím bakalářské práce.

Datum	Podpis

Poděkování

Rád bych poděkoval vedoucímu práce za cenné rady pro úspěšné dokončení práce.

Abstrakt

Tato bakalářská práce se zabývá uplatněním statistických modelů k predikci umístění závodníků v biatlonu. Práce je také navržena k vytvoření databázové struktury, která dále usnadní přístup ke stávajícím statistikám a sama tyto informace aktivně využívá pro modelování výsledků. Statistické modely jsou obohaceny o dynamické rozšíření, které je využito k výpočtům aktuální výkonnosti závodníků. Práce dále může sloužit díky obecnému zápisu funkcí jako program pro vytváření nových biatlonových modelů při zachování základní logiky aplikace. Práce v neposlední řadě prezentuje své dosažené hodnoty pomocí webové aplikace. Důraz je kladen na zabezpečení aplikace, možnost modifikace a dostatečné rychlosti i při vysokém zatížení.

Klíčová slova: biatlon, sportovní analýza, statistické modely

Abstract

The thesis is devoted to application of statistical models to predict the placing of athletes in biathlon. The work is also designed to create adatabase structure that will facilitate access to existing statistics, and it uses this information for modeling results. Statistical models are enriched with a dynamic extension, which is used to calculate the actual performance of athletes. Thanks to the general record of functions, the work may also be used as a program for creating new biathlon models while maintaining the basic logic of the application. Last but not least, the work presents its values achieved using a web application. Emphasis is placed on application security, the possibility of modification andsufficient speed even at high loads.

Keywords: biathlon, sports analysis, statistical models

Seznam tabulek

Obsah

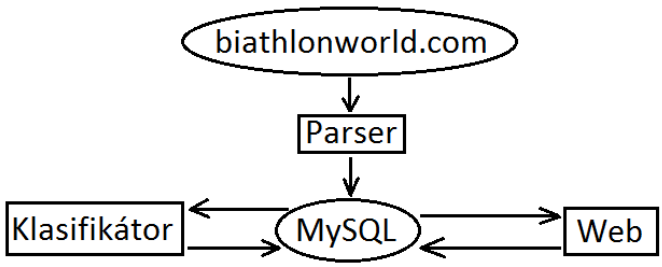
1 Úvod

Již několik desítek let je drtivá většina sportovních odvětví ve světové špičce ryze profesionální záležitostí. Jedním z faktorů, který stojí za touto profesionalizací, je sázení na sportovní příležitosti. Sázkové kanceláře, které tento populární druh zábavy poskytují, patří mezi významné firmy. Jejich roční obrat na evropském trhu pouze na online sportovní sázení činil v roce 2010 3,4 mld. USD, celkový obrat sázení (včetně pokeru a dalších loterií) dosahuje až 350 mld USD [1]. Největší on-line sportovní sázková kancelář The bwin Group překročila celkový počet 20 milionů registrovaných zákazníků [2].

Mezi nejpopulárnější sporty v Evropě, na které jsou uzavírané sázky, patří fotbal, lední hokej, basketbal a také dostihy, které mají zejména v Anglii velkou tradici. Biatlon se mezi tyto velikány sice v současné době zařadit nemůže, ovšem v posledních letech zažívá velký rozmach. Tento trend zachytila i většina sázkových kanceláří, a tak si u většiny z nich je možné vsadit na nejrůznější výsledky biatlonu. Vzhledem k vysokému obratu sázkových kanceláří lze i s menší popularitou biatlonu usuzovat, že si i tento sport zaslouží patřičnou pozornost a může, jak již dnes vypisováním kurzů sázkové kanceláře dokazují, být i ekonomicky výnosný.

Tato práce připravuje systém pro modelování biatlonových výsledků a také návrh jejich vyhodnocení. Rovněž navrhuje i konkrétní řešení.

Veškeré výpočty a logika aplikace bude naprogramována v jazyku Java, statistiky a výsledky ukládány do databáze MySql. Tato databáze sdílí data jak s výpočetním řešením, tak i s webovou aplikací, která poslouží k prezentaci získaných výsledků a vykreslení statistik. Základní funkcionalita programu je zobrazena na Obrázku 1.



Obrázek 1: Návrh aplikace

Ve 2. kapitole se provádí analýza biatlonu a jeho zákoutí, jsou zde vysvětleny rozdíly mezi jednotlivými druhy závodů, zmíněny aktuální trendy a nastíněny nejdůležitější faktory, ovlivňující výsledky tohoto sportu.

Kapitola 3 pojednává o biatlonových statistikách a možnostech, kde tato data získat. Jelikož všechny informace nemusí být k dostání a také mohou obsahovat chyby, je v této kapitole vyhodnocena i průměrná chybovost jednotlivých dat.

Kapitola 4 představuje statistické modely a jejich současný stav řešení i v ostatních sportech [3]. Dále je zde nastíněna struktura modelů, která by měla být dodržena pro všechny dále vyvíjené modely aplikace, aby byla zachována dostatečná obecnost řešení k rychlému navázání na vývoj této práce. Z tohoto důvodu je nutné dopředu stanovit výstupy modelů, které následně mohou být vyhodnocovány pomocí jedné funkcionality.

Kapitola 5 uvede model založený na předcházejících výsledcích závodníků. Pomocí dynamické úpravy a zavedení vah umístění se vyhodnocuje aktuální forma závodníků. Model je také rozšířen o redukci nepovedených výsledků. Dále je navržena tabulka vztahů mezi typy závodů, pomocí níž je pak možné přidělovat vyšší prioritu předem vybraným typům závodů.

V 6. kapitole je představen nový model, který vzniká modifikací předchozího. Místo umístění jsou výsledky skládány z časových rozložení běhu a střelby.

Kapitola 7 vyhodnocuje dosažené výsledky a porovnává úspěšnost statistických modelů.

8. kapitola uvede webovou aplikaci sloužící k prezentování statistik a predikovaných výsledků.

V posledním kapitole se dozvíme přínosy této práce a možnost jejího dalšího rozvoje, vedoucího ke zlepšení modelů.

2 Biatlon

Biatlon je zimní sport, který kombinuje běžecké lyžování se střelbou. Počátky tohoto sportu v moderním pojetí sahají do roku 1955, kdy byla sepsána první pravidla. Od roku 1982 pak začaly na mezinárodní úrovni závodit i ženy.

2.1 Základní pojmy

Střelecká položka se skládá z pěti střeleckých pokusů. Každá položka v závodě má striktně určeno, jestli bude odstřílena ve stoje, či v leže. Terče jsou umístěny ve vzdálenosti 50 metrů a jsou tvořeny kruhovým tvarem o rozměrech 110mm pro položku ve stoje, resp. 45 mm pro položku v leže. Za každou chybu musí závodník absolvovat trestný okruh, který měří 150 metrů, případně u vytrvalostního závodu je mu k výslednému času přičtena trestná minuta.

2.2 Typy závodů

Vytrvalostní závod i sprint startují závodníci s rozestupy 30 sekund, startovní čísla se losují, avšak jsou rozdělena do 2-4 skupin dle počtu startujících. Prvních 10 závodníků průběžného hodnocení má právo být nasazeno v libovolné skupině [4]. Do stíhacího závodu se kvalifikuje prvních 60 závodníků ze sprintu a následně vybíhají Gundersenovou metodou dle výsledků sprintu. Závodu s hromadným startem se účastní 30 závodníků, kteří vybíhají ve stejný čas.

Další důležité parametry jsou znázorněny v Tabulce 1.

Typ závodu	Běh - ženy	Běh - muži	Střelba	1 střelecká chyba
Vytrvalostní	15 km	20 km	L,S,L,S	60 s k výslednému času
Sprint	7,5 km	10 km	L,S	1 trestný okruh
Stíhací závod	10 km	12,5 km	L,L,S,S	1 trestný okruh
Závod s hromadným startem	12,5 km	15 km	L,L,S,S	1 trestný okruh

Tabulka 1: Porovnání typů individuálních závodů

L,S = první položka v leže (L), druhá ve stoje (S)

2.3 Faktory ovlivňující závod

Biatlon se skládá ze dvou rozdílných disciplín, běhu a střelby. Je tedy možné oddělit čas, který závodník stráví střelbou a během, ačkoliv tyto parametry na sobě také nepřímo závisí.

2.3.1 Rychlost běhu

Běh se považuje za nejdůležitější součást biatlonu, bez jeho solidního zvládnutí je šance na solidní umístění prakticky vyloučena. Běžecká výkonnost nejvíce závisí na celkové fyzické připravenosti závodníka. Velmi důležité je však i zvládnutí samotné běžecké techniky, délka skluzu běžce může značně zefektivnit vynaloženou námahu, zejména pak v rovinatých pasážích. Dalším faktorem, který může pomoci, je stabilita, která pomáhá v těžkých pasážích, především ve sjezdech.

Konkrétní běžecký okruh sehrává také svou roli, velké převýšení svědčí spíše lehčím sportovcům, kteří překonávají menší gravitační sílu. Stejně tak se obecně pokládá za pravdu, že lehčím závodníkům kromě kopcovitějších tratí svědčí také delší trasy, tento faktor však není tak podstatný. Nadmořská výška v místě konání se rovněž podepíše na výkonech.

Aktuální podmínky jsou další zajímavou součástí běžeckého výkonu. Teplota vzduchu ovlivňuje energetickou spotřebu závodníků a jejich potřebu dodržovat pitný režim, dále však přímo ovlivňuje teplotu sněhu. Zmrzlý sníh poskytuje určitou výhodu technicky dobře vybaveným běžcům, hlubší (měkký) sníh pak lehčím závodníkům. Nejdůležitější roli a překvapivé zvraty pak mohou nachystat sněhové podmínky, pokud v průběhu závodu začne, či přestane sněžit trať se začne zpomalovat resp. zrychlovat. Tento aspekt se projeví při sprintu a vytrvalostním závodu, kdy závodníci startují s poměrně velkými rozdíly a měnící se sněhové podmínky mohou ovlivnit výsledky závodu. Možnost výše zmíněného nasazení závodníka do losovaných skupin může národní federace pomocí svého trenéra výrazně ovlivnit, při možnosti měnícího se počasí pak mohou trenéři strategickou volbou ovlivnit výsledky svých závodníků.

V současném startovním poli vyčnívá svými běžeckými schopnostmi Martin Fourcade, vítěz světového poháru v sezóně 2012/2013. Dalším zajímavým závodníkem, který ovládá běžecké tratě je Emil Hegle Svendsen, který však vyniká také výborným závěrem, což mu poskytuje výhodu při hromadném i stíhacím závodu, tuto sezónu tímto způsobem zvítězil např. ve stíhacím závodu na Mistrovství světa v Novém Městě na Moravě. V ženském pelotonu v současné době udává běžecké časy Miriam Gössner, 4. z mistrovství světa v klasickém lyžování.

2.3.2 Přesnost střelby

Přesnost střelby je z hlediska vývoje závodu velmi důležitým bodem. Většina závodníků má vyšší úspěšnost při položce v leže.

Střeleckou úspěšnost mimo samotného natrénování výrazně ovlivňuje počasí, zejména měnící se směr větru. Často jsme z tohoto důvodu během závodu svědky přenastavování mířidel, které se provádí těsně před samotnou střelbou, závodník touto úpravou neztrácí prakticky žádný čas. Pro závodníky je při absolvování položky rovněž nepříjemné husté sněžení, či dokonce déšť. Často může rozhodnout i psychická odolnost, dokonce býváme i svědky odstřílení kompletní položky do terčů soupeře, tuto obrovskou chybu už si dokázali připsat i nejlepší biatlonisté např. Magdalena Neunerová či Andreas Birnbacher.

Nejlepší střelkyní uplynulé sezóny se stala Marie Laure Brunet s úspěšností 93,3%, předčila tak dokonce i všechny mužské závodníky.

2.3.3 Rychlost střelby

Rychlost střelby představuje celkový střelecký čas včetně zaujetí postoje až do opuštění střeleckého stavu. Význam tohoto parametru je oproti výše uvedeným nejmenší, přesto při vyrovnanosti startovního pole je možné i zde rozhodnout o výsledku závodu.

Za jednoho z průkopníků detailního zvládnutí rychlosti střelby, včetně nejrychleji možného zaujetí střeleckého postoje, můžeme považovat legendu biatlonu Ole Einara Bjoerndalena. Jeho velmi zdatným nástupcem se stává Simon Eder, který je schopen kompletní střeleckou položku zvládnout do 20 s.

3 Statistiky

Pro výpočet pravděpodobností sportovních událostí je nutné získat co nejkomplexnější výsledky z minulosti. S rostoucím počtem získaných informací se zvyšuje i počet trénovacích a testovacích dat, což vede ke zlepšení úspěšnosti klasifikace.

3.1 Biatlonové statistiky

Jako stěžejní data, která budou potřeba k vytvoření statistik byly určeny:

- v každém okruhu počet střeleckých chyb, čas střelby a celkový čas okruhu
- umístění závodníků v cíli

Potřebná data jsou dostupná z oficiálního serveru biathlonresults.com. Všechny informace lze získávat pouze z pdf souborů, které jsou přiděleny k jednotlivým závodům.

3.2 Získání a uložení dat

Pro uchování získaných dat byla vybrána databáze MySql, získávání dat ze souborů ve formátu pdf probíhá s pomocí programovacího jazyka Java a knihovny na čtení těchto souborů PDFTextStream, který je stejně jako výše uvedené nástroje s omezením víceúlohových aplikací dostupný zdarma. Známější opensource knihovna iText nesplnila požadavky, jelikož často nedokázala rozeznat textový výstup, který se tak následně sléval do jednoho slova bez oddělení. Další problém představuje samotný formát souborů, který není úplně striktní, často ani kompletní. Zejména u závodů druhé kategorie tzv. Ibu-Cupu nejsou mnohdy k dispozici údaje o střelbě, což následně mění i formát PDF souboru a jeho práci s ním.

Každou sezónu se koná 26 závodů nejvyšší úrovně (mistrovství světa, olympijské hry + světový pohár) jak u mužů, tak i u žen. Pro závodníky, kteří se nedokáží měřit se světovou špičkou, je určen Ibu-cup, který je nástupcem evropského poháru, každoročně se také pořádá mistrovství jednotlivých kontinentů. Vzhledem k tradicím a rozšířenosti biatlonu kvalitativně evropský šampionát výrazně převyšuje ostatní, přesto se těchto mistrovství většina světové špičky neúčastní. Posledním typem seniorských závodů jsou národní šampionáty, které se často využívají ke kvalifikaci do národního týmu.

3.3 Chybovost dat

Jelikož záznamy o závodech jsou často nekompletní nebo obsahují chybné informace, je nutné stanovit míru chybovost dat. Vyloučit nemůžeme ani chybu našeho parseru, zvláště pokud pracuje s dalšími knihovnami, nicméně pravděpodobnost chyby můžeme pokládat za téměř nulovou.

3.3.1 Výpočet chybovosti

V případě, že program na stahování dat nenalezl požadovanou hodnotu, vrací NULL, tato hodnota se zaznamenává do databáze a dle četnosti tohoto příznaku můžeme určit počet střeleckých položek, kde nebyla zaznamenána žádná hodnota. Pokud by při jediné položce byl počet střeleckých chyb větší než 5, pak se jedná také o chybu zanesených dat.

U střeleckých časů vyhodnocujeme data obdobným postupem, pouze u určení chyb z vlastních zkušeností budeme považovat rychlejší střelbu než 15s za chybně zapsanou.

Běžecké časy stejně jako předchozí parametry mají nedefinovanou hodnotu, pokud nebyla zanešena. Jelikož běžecký okruh nemůže měřit méně než 2,5km, můžeme i časy pod 3 minuty rovněž považovat za chybné.

Úspěšnost (doplněk k chybovosti) jednoduše vyjádříme jako:

$p = 1 - \frac{k}{n}$;n – celkový počet dat; k – celkový počet chybných dat

	Položky - celkem	Chyby (NULL)	Počet chyb > 5	Celkem chyb	Úspěšnost
WrlCp, Wrlch	33264	121	0	121	99,64%
Ostatní	22394	1190	0	1190	94,69%
Celkem	55658	1311	0	1311	97,64%

Tabulka 2: Chybovost střeleckých položek

	Položky - celkem	Chyby (NULL)	Střelba < 15s	Celkem chyb	Úspěšnost
WrlCp, Wrlch	33264	121	9	130	99,61%
Ostatní	22394	9144	9	9153	59,13%
Celkem	55658	9265	18	9283	83,32%

Tabulka 3: Chybovost střeleckých časů

	Položky - celkem	Chyby (NULL)	Čas < 3min	Celkem chyb	Úspěšnost
Celkem	74743	0	0	0	100,00%

Tabulka 4: Chybovost časů

3.3.2 Vyhodnocení chybovosti dat

Z jednoduchého testu dat dojdeme k závěru, že u významějších závodů oficiální web poskytuje kompletnější výsledky. Tabulka 3 ukazuje velkou nekompletnost střeleckých časů u méně důležitých závodů, použití takovýchto údajů se může ukázat jako kontraproduktivní.

Ačkoliv chybovost časů je nulová, pro přesné určení jednotlivých časů jsou nutné k těmto datům i kompletně zaznamenané údaje o střelbě. Běžecké časy jsou tedy závislé na střelbách. Výsledná úspěšnost je tedy shodná s úspěšností střeleckých časů.

4 Statistické modely

4.1 Známá řešení

Existující řešení modelů jsou navrhována pro nejsledovanější sporty jakými jsou fotbal, hokej, basketbal, americký fotbal a mnohé další. Tyto sporty však mají kromě své popularity společnou i další vlastnost. Jedná se o kolektivní sport, soupeří proti sobě pouze dva týmy, existují 3 možné výsledky (výhra, remíza, prohra).

Tyto sporty se např. u fotbalu a hokeje vyhodnocují na základě výsledků, který rozlišuje pouze výhry, remízy, prohry. Další možnost představuje počet vstřelených branek, obvykle se vyhodnocuje i výhoda domácího prostředí [3].

Sázkové kanceláře však své informace důkladně tají, není tak znám aktuální stav jejich řešení, nepotvrzují ani existenci statistických modelů, které však pro známější druhy sportů, než biatlon, jistě využívají.

4.2 Použitá data pro klasifikaci

Na základě vyhodnocení chybovosti dat a pečlivého zvážení důsledků zanedbání některých údajů byly vybrány pro účel klasifikace pouze výsledky ze světových pohárů a mistrovství světa. Závody dalších úrovní nebudou brány v úvahu, jelikož se jich účastní světová elita jen velmi zřídka, chybovost získaných údajů je podstatně vyšší a zejména pak riziko, že z těchto soutěží (Ibu-cup) pronikne do světového poháru závodník, který začne od prvních závodů výrazně promlouvat do konečného umístění na předních pozicích, se považuje za zanedbatelné. Ostatní závody již nejsou při klasifikaci použity, jelikož jejich kvalita již zdaleka nedosahuje potřebné úrovně.

Použitý soubor dat za sezóny 2010/2011, 2011/2012, 2012/2013 ze světových pohárů a mistrovství světa poskytuje 156 závodů. Ve srovnání s vyhodnocováním fotbalových modelů by se mohl tento rozsah zdát nedostatečný. Klasifikace ovšem probíhá pro každého závodníka v každém ze závodů, klasifikovaných objektů je tedy 11 168.

4.3 Způsob vyhodnocení modelů

4.3.1 Trénovací data

Každý model, který byl vytvořen, přebírá trénovací data, tj. soubor všech potřebných údajů pro predikci výsledků v zadaném závodu.

Jelikož je nutné, aby vyhodnocení úspěnosti modelů odpovídala co nejvíce reálným hodnotám, byly výsledky z celé sezóny 2010/2011 zařazeny do trénovací množiny, čímž se sníží i počet klasifikovaných objektů na 7 530. S každým dalším vyhodnocením závodu se trénovací data rozrostou o jeden závod. V trénovací množině tak jsou k dispozici neustále všechny data, která by byla dostupná i před závodem, který je aktuálně vyhodnocován.

4.3.2 Testovací data

Testovací data v modelech představují závodníky, kteří nastoupí do závodu, na kterém je prováděna klasifikace. Startovní listina je k dispozici před každým závodem na oficiálním serveru biathlonworld.com, která obsahuje seznam závodnických ibuid (identifikátor každého biatlonisty určený jeho národností, datumem narození, pohlavím, v případě potřeby dalšími údaji). Každý závod se dále identifikuje pomocí datumu, kategorie, pohlaví a typu závodu.

4.3.3 Vyhodnocení modelů

Každým modelem jsou stanoveny výsledky jednotlivých závodníků v cíli. Tímto způsobem by ovšem nebylo možné stanovit konkrétní pravděpodobnosti sledovaných jevů. Dalším požovaným výstupem modelu tedy musí být také struktura, která pro každého závodníka určí pravděpodobnost všech možných umístění.

Výsledné hodnoty úspěšnosti klasifikace se stanovují na základě všech klasifikovaných závodů. V každém závodu nejdříve stanovíme průměrnou odchylku ϕ podle vzorce (4.1a), kde n značí počet závodníků, x_i predikovaný výsledek a y_i skutečný výsledek i-tého závodníka.

$$\phi = \frac{1}{n}\sum_{i=1}^{n} |x_i - y_i| \quad (4.1a)$$

K docílení vyšší vypovídající hodnoty odchylky umístění stanovíme relativní úspěšnost.

$$\phi_p = 1 - \frac{1}{n^2}\sum_{i=1}^{n} |x_i - y_i| \quad (4.1)$$

Dále stanovíme čtvercovou odchylku, která slouží ke zvýraznění odchylek, jelikož vyšším odchylkám přisuzuje vyšší váhu.

$$\phi^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \quad (4.2)$$

Výše uvedené odchylky ϕ, ϕ^2 v ideálním případě nabývají hodnoty 0. Nižší odchylka pak tedy znamená lepší výsledky statistického modelu.

Maximální odchylka od skutečného umístění se rovněž často uplatňuje, nicméně již je částečně obsažena ve čtvercové odchylce. Této hodnotě rovněž není přisouzen velký význam zejména z důvodů, že se jedná o velmi nestálou hodnotu a velká odchylka jednoho závodníka může být způsobena různými nepředvídatelnými okolnostmi, které je však rovněž nutno brát na vědomí.

Posledním parametrem, kterým se stanoví úspěšnost schopnosti klasifikace modelu, je součet pravděpodobnostních umístění (spi) všech závodníků. Pravděpodobnostní umístění je modelem určená pravděpodobnost, odečtená ze skutečného výsledku závodu. Výsledná hodnota tedy může nabývat rozmezí 0 – n, kde n představuje počet startujících. Pokud by tedy stanovená pravděpodobnost pro všechny závodníky byla konstantní, tj. 1/n, součet pravděpodobnostních umístěni by se rovnal jedné. Se zvyšujícím se počtem závodníků se tedy teoreticky zvedá i tato hodnota, musíme však mít na paměti, že se zvyšuje i počet možností, na které se pravděpodobnost rozkládá. Tato suma by se nikdy se zvyšujícím počtem závodů v teoretické množině dat neměla nikdy snižovat, může pouze zůstat na stejné hodnotě (1), v případě, že všichni startující mají stejnou šanci na vítězství, i na ostatní pozice.Všechny parametry jsou vyhodnoceny pro každý klasifikovaný závod, spočteme tedy celkové výsledky klasifikace jako aritmetický průměr (4.3).

$$E(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (4.3)$$

Samotné vyhodnocení se provádí na celé množině dat. V některých případech se mohou použít modely pouze pro specifické podmínky, často se tedy používají určité podmnožiny klasifikace. V této práci byly vybrány jednotlivé druhy závodů a zvláště jsou také vyhodnoceny úspěšnosti klasifikace u žen a mužů.

5 Poziční model

Jediným příznakem, kterým se stanovují výsledky u tohoto modelu, jsou umístění v předcházejících závodech. Predikce výsledků je stanovena pomocí aritmetického průměru, který představuje předpokládané umístění závodníka v klasifikovaném závodu.

5.1 Redukce umístění

Tento způsob predikce ovšem nebere v úvahu počet startujících, což může způsobit, že předpovídané pozice budou vyšší než počet závodníků.

Zmíněný problém se řeší stanovením maxima, kterého může nabývat predikovaná hodnota. Toto maximum odpovídá počtu účastníků, kteří zasáhnou do závodu. U závodníků, kterým bude pomocí této redukce změněno předpokládané umístění se ovšem dopouštíme vůči ostatním závodníkům zlepšení jejich výsledku bez zřejmého důvodu. Mírnou modifikací se určí koeficient (5.1), kde m představuje největší průměrnou hodnotu, která byla spočtena, n pak celkový počet startujících v klasifikovaném závodu.

$$c=(m-1)/(n-1);n\geq 2 \quad (5.1)$$

Dosazením do (5.2) získáme výsledné umístění pro i-tého závodníka, y_i představuje pozici v cíli vytvořenou modelem.

$$x_i=\frac{(y_i-1)}{c}+1 \quad (5.2)$$

Každým zásahem do vyhodnoceného výsledku se pravděpodobnost mírně změní, tudíž v zásadě jsou tyto metody nepoužitelné v případě, že je nutné rozložení pravděpodobnosti bezpodmínečně zachovat. Pokud by se však redukce nepoužila, výsledky by rovněž nebyly příliš vypovídající, ačkoliv zachování rozložení by bylo splněno, výsledné umístění by až příliš ovlivnil počet startujících, proto ani tato možnost k optimálnímu řešení nevede.

Hlavním úkolem při predikci výsledků však není stanovit nejpravděpodobnější výsledek, ale pravděpodobnostní rozložení všech možných umístění. Rozložení by se stanovilo pomocí funkce, která na základě rozdílů vypočtených redukovaných hodnot určuje pravděpodobnosti, kde větší rozdíl znamená vyšší pravděpodobnost dobrého výsledku pro závodníka s menším průměrem umístění. Tento výpočet by však byl velmi nepřesný, jelikož zpětně z průměrů získává rozložení, které lze stanovit přímo ze vstupních veličin. Dále tedy aritmetický průměr minulých umístění

závodníka je považován za odhad jeho výsledné pozice v cíli. Aby tento odhad byl reálný je redukován prostým vzestupným seřazením, kde nové hodnoty jsou přepsány a nabývají hodnot 1 až n = počet startujících.

5.2 Pravděpodobnostní rozložení

Pravděpodobností rozložení je vyjádřeno jako normální (Gaussovo) rozdělení (5.3).

$$f(x)=\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-E(x))^2}{2\cdot\sigma^2}} \quad (5.3)$$

K jeho výpočtu se nejrpve stanoví rozptyl vzorcem (5.4), kde n představuje počet všech umístění v předcházejících závodech, x_i konkrétní umístění závodníka v i-tém závodu.

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}[x_i-E(x)]^2 \quad (5.4)$$

E(x) vyjadřuje aritmetický průměr všech umístění (4.3).

Gaussovo rozdělení v původní podobě nelze využít, jelikož připouští rozmístění i pro záporná umístění, která jsou nežádoucí, jelikož předpokládáme nejlepší možné umístění jako první místo. Záporná umístění ve své podstatě mohou mít svůj význam, předpokládají lepší výsledek, nežli umístění nejlepšího závodníka, což v určení pravděpodobnosti jistě může nastat. Příliš velký rozptyl umístění závodníků by ovšem přinesl příliš velkou pravděpodobnost na přední umístění, ačkoliv jejich závodní výsledky nemají potenciál reálně proniknout mezi nejužší světovou špičku. Rozdělení tak uvažujeme jen od počáteční hodnoty (1. místo) až do n, kde n ovšem nepředstavuje celkový počet, ale dostatečně velkou hodnotu, po které již součet pravděpodobností je dostatečně nízký, než aby ovlivnil další výpočty a zkreslil tak jejich hodnoty. Celkový obsah S (součet pravděpodobností) by se stanovil jako integrace vybraného úseku, viz vzorec (5.5), jelikož však tuto integraci normálního rozdělení nelze vyjádřit pomocí elementární funkce, musíme využít méně přesné numerické integrace.

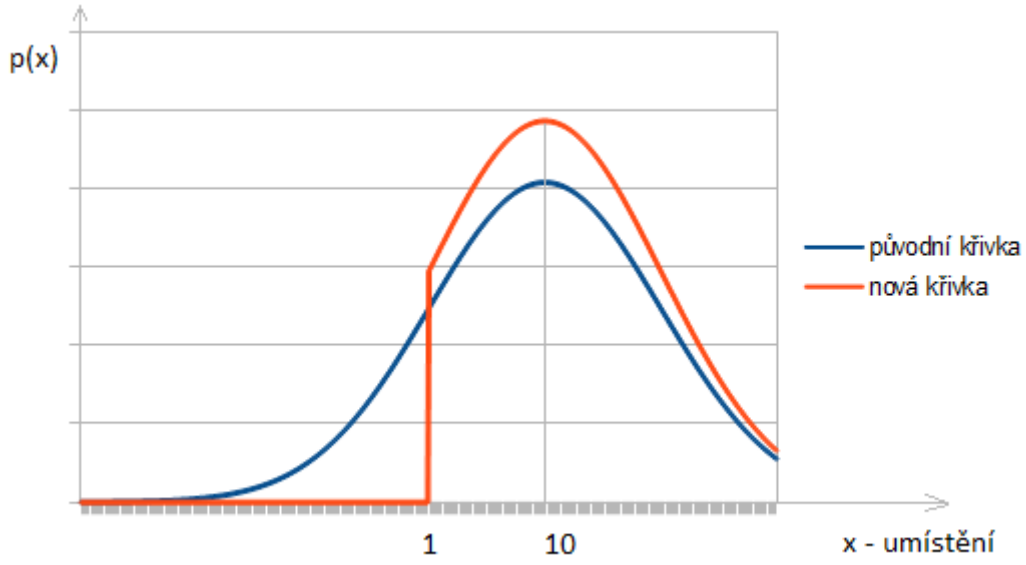
$$S=\int_{1}^{n}\frac{1}{\sigma\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(t-E(x))^2}{2\cdot\sigma^2}}\,dt \quad (5.5)$$

Hodnota numerické integrace se vypočte dosazením do vzorce (5.6), kde parametry a, n představují zkoumaný interval, d pak rozdělení jednotky na počet dílů. V ideálním případě se počet

intervalů, které jsou sčítány blíží nekonečnu.

$$o = \sum_{i=a}^{n} \sum_{j=0}^{d-1} \frac{\frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{\left((i+\frac{j}{d}) - E(x)\right)^2}{2 \cdot \sigma^2}}}{d} \quad (5.6)$$

Spočtením numerické integrace ze vzorce (5.6) na intervalu od prvního místa (a = 1) do n-tého umístění, které bylo zvolena jako b = 250, jelikož nejvyšší počet startujících v některých závodech se pohybuje i výrazně přes 100. Počet dělení jednotky d se stanovil na 100, což celkem představuje rozdělení funkce na 25 000 intervalů, což zajistí dostatečnou přesnost výpočtu a zachová i jeho rychlost. Tato hodnota se invertuje a vynásobí se jí celý zjištěný interval, který se však také vyjadřuje jako obsah jednotlivých intervalů.V Obrázku 2 je znázorněno Gaussovo rozložení po transformaci do určeného intervalu.



Opakováním zmíněného výpočtu pro všechny závodníky bylo zjištěno pravděpodobnostní rozložení závodníků pro konkrétní umístění. Výsledná pravděpodobnost na zadanou pozici pak je stanovena jako součet všech obsahů mezi vybranými intervaly.

5.2.1 Redukce pravděpodobnostního rozložení

Pravděpodobnostní rozložení ponechané ve stavu, který je výše představen, může poskytovat nejlepší možné výsledky, jen pokud není stanoven počet startujících závodníků a není znám ani žádný další parametr, který by upřesnil šance závodníka. Takovým parametrem by mohl být údaj o typu závodu nebo i další informace, které by zúžily startovní listinu, nebo by ji pomohly určit.

Ideální variantu redukce pravděpodobnostního rozložení představuje metoda založená na součtu křivek pravděpodobnostího rozložení všech závodníků a následné rozdělení jejich pravděpodobností do jednotlivých umístění.

Algoritmus řešící toto vyhodnocení nejdříve přejme vygenerované pravděpodobnostní rozložení (p) pro všechny závodníky (celkový počet závodníků n) na startovní listině, v každém rozložení je obsažen stejný počet hodnot, u námi zvoleného rozdělení a intervalu je počet 25 000. Množina obsahuje seřazený soubor dat od pravděpodobností na nejlepší výsledky až po nejhorší, přičemž u všech závodníků je struktura identická. Výpočetní funkce alokuje n nových datových struktur, do kterých se v průběhu výpočtu vkládají pravděpodobnosti určené ke konkrétní hodnotě. Výpočet probíhá neustále v cyklu, dokud nevyhodnotí všechny možnosti umístění.

Výhodou tohoto algormu je jeho univerzálnost, pracuje sice s konkrétním souborem dat, což může představovat horší úspěšnost v závislosti na počtu intervalů, než u matematicky zapsané funkce. Pokud jsou zachovány podmínky pro tento druh redukce umístění ,je možné jej nasadit na libovolnou funkci, nejen normální rozdělení.

Poziční klasifikátor založený pouze na aritmetickém průměru předchozích umístění závodníka dosahuje po klasifikaci výsledků, které jsou znázorněny v Tabulce 5. Detailní vyhodnocení modelů je popsáno v kapitole 7.

Množina dat	ϕ_p	ϕ^2	spi	Max. odchylka
ženy	82,58%	20911,1	1,8880	60,3300
muži	81,01%	29426,9	1,7371	65,8000
Vytrvalostní závod	83,69%	48314,6	2,0090	60,3100
Hromadný start	76,46%	2441,8	1,3068	66,9500
Stíhací závod	82,83%	8423,84	1,7546	58,7900
Sprint	83,08%	42985,1	2,0529	65,3700
Celkem	81,80%	25169	1,8125	63,0700

Tabulka 5: Vyhodnocení úspěšnosti pozičního modelu v závislosti na aritmetickém průměru

5.3 Dynamické rozšíření modelu

Závodníkova výkonnost v průběhu sezóny se mění, ať již v závislosti na zdravotním stavu, načasování tréninku či jiných okolnostech. Tento jev bývá označován jako forma. K aktuálnímu příznaku umístění v předcházejících závodech přibývá další veličina, která bude dále označována za váhu (důležitost) tohoto příznaku. Se zavedením dalšího parametru se změní i způsob výpočtů. Aritmetický průměr (4.3) se modifikuje na vážený s přidáním váhy w_i i-tého umístění. Průměr se pak určí dle vzorce (5.7).

$$E(x)=\frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i} \quad (5.7)$$

Stejně tak i rozptyl (5.4) je zaměněn za vážený rozptyl (5.8), kde E(x) představuje vážený průměr.

$$\sigma^2=\frac{\sum_{i=1}^{n} (y_i - E(x))^2 \cdot w_i}{\sum_{i=1}^{n} w_i} \quad (5.8)$$

Vážený rozptyl a vážený průměr je následně dosazen do vzorce (5.6)

Forma závodníků se v čase mění a tak je odvozena z datumu závodů. Ve výpočetní aplikaci je definována jako libovolná funkce, která jako parametr přejímá celé číslo, (d_f) představující rozdíl vyjádřený počtem dní mezi klasifikovaným a trénovaným závodem. Váha pak přímo závisí na formě a může být vyjádřena např. zápisem w = (730 - d_f). Váha podléhá podmínce, že nikdy nemůže být nižší než 0, pokud by se tak stalo, automaticky je nastavena na nulu a příznak tak nebude mít žádný vliv na predikci. Výše uvedený zápis by tedy zanedbal všechny závody konané před více než dvěma roky a nejvyšší možnou váhu (730) by měl závod konaný v den samotného klasifikovaného závodu. Ačkoliv by se mohlo předpokládat, že váha může nabývat pouze hodnot <0,1> není to nutné, jelikož se vypočtená hodnota dle vzorců (5.7,5.8) dělí sumou vah.

Množina dat	ϕ_p	ϕ^2	spi	Max. odchylka
ženy	82,95%	20229,6	1,9263	59,7200
muži	81,29%	28854,83	1,7773	65,2800
Vytrvalostní závod	83,88%	47762,25	2,0663	60,1300
Hromadný start	76,75%	2359,75	1,3207	66,6100
Stíhací závod	83,32%	7886,94	1,7931	57,0900
Sprint	83,31%	41991,65	2,1000	65,4800
Celkem	82,12%	24542,21	1,8518	62,5000

Tabulka 6: Vyhodnocení úspěšnosti pozičního modelu se zahrnutím formy

Funkce váhy w = (730 - d_f) nedostatečně popisuje aktuální výkonnost, nedostatečně upřednostňuje nedávné výsledky, proto vyzkoušíme úspěšnost klasifikace pomocí další funkce

$$w = \frac{1}{\sqrt{(d_f)}}$$

Množina dat	ϕ_p	ϕ^2	spi	Max. odchylka
ženy	83,79%	19477,04	1,9593	58,32%
muži	82,12%	28365,15	1,7963	64,34%
Vytrvalostní závod	83,80%	47618,17	2,0776	59,91%
Hromadný start	77,50%	2227,25	1,3209	67,10%
Stíhací závod	85,46%	6297,19	1,8683	53,90%
Sprint	83,42%	41758,03	2,1040	64,83%
Celkem	82,95%	23921,1	1,8778	61,33%

Tabulka 7: Vyhodnocení pozičního modelu se zahrnutím vylepšeného parametru formy

5.4 Redukce náhodnosti umístění

Závodník během sezóny absolvuje mnoho závodů a tak se mu některé nemusí úplně povést a výrazně by pak snižovaly jeho průměrný výkon. Tento faktor je nutné redukovat zejména z důvodu, že závodník, kterému se během závodu nedaří a nemá reálnou šanci na solidní umístění více riskuje při střelbě, nebo naopak poslední běžecká kola již jede nižším tempem, než obvykle.

Tento faktor se vyjádří pomocí váhy. Jelikož do formy vstoupí další parametr, celková váha se tedy vypočte jako násobek všech dílčích vah, které obsahuje. Samotná váha redukovaného umístění se vypočte sestupným seřazením všech (n) výsledků závodníka. Každé hodnotě se přidělí index i. U nejvyššího umístění se i=0. U každé další hodnoty se index iteruje tak, že u nejvyšší hodnoty je index i roven n – 1. Dále se spočte parametr d_i pro každou hodnotu dle vzorce (5.9).

$$d_i = \frac{i}{n-1} \quad (5.9)$$

Pomocí tohoto parametru se opět docílilo obecnosti řešení a nyní můžeme zvolit váhu např. Jako $w = \sqrt{(d_i)}$. Touto funkcí je možné také redukovat nejlepší umístění závodníka. Stále se totiž pracuje s parametrem d_i , a tak je možné přepsat redukční funkci umístění jako $w = \sqrt{(1 - d_i)}$

Množina dat	ϕ_p	ϕ^2	spi	Max. odchylka
ženy	83,78%	19522,5	2,0405	58,46%
muži	82,23%	28171	1,9085	63,86%
Vytrvalostní závod	83,90%	47339	2,1357	59,16%
Hromadný start	77,40%	2258,65	1,3659	66,93%
Stíhací závod	85,63%	6255,63	1,9920	54,01%
Sprint	83,44%	41666,03	2,2165	64,60%
Celkem	83,00%	23846,75	1,9745	61,16%

Tabulka 8: Vyhodnocení pozičního modelu se zahrnutím formy a redukce umístění

5.5 Závislost výsledků na typech závodů

V biatlonu existují 4 individuální disciplíny, viz. 2. kapitola. Všechny tyto závody se od sebe mírně odlišují a tak lze předpokládat, že v každém ze závodů mají závodníci rozdílné šance na úspěch. Tato problematika je řešena stanovením tabulky závislostí, kde jsou vyjádřeny všechny vztahy mezi závody, pomocí přidělení významu (vah) vybraným závodům.

Tato tabulka může vypadat následujícím způsobem:

	Vytrvalostní závod	Sprint	Stíhací závod	Hromadný závod
Vytrvalostní závod	1	0.75	0.5	0.5
Sprint	0.75	1	0.75	0.5
Stíhací závod	0.75	1	1	1
Hromadný závod	0.75	0.75	0.75	1

Tabulka 9: Závislosti vah závodů

Tabulka 9 znázorňuje ve sloupci klasifikovaný závod a přiřazuje mu váhy pro známé typy závodů.

Množina dat	ϕ_p	ϕ^2	spi	Max. odchylka
ženy	83,85%	19459,88	2,0421	58,19%
muži	82,26%	28140,96	1,9094	64,10%
Vytrvalostní závod	83,94%	47197,5	2,1213	59,00%
Hromadný start	77,55%	2250,65	1,3662	67,60%
Stíhací závod	85,66%	6202,75	2,0058	53,50%
Sprint	83,47%	41634,33	2,2128	64,68%
Celkem	83,06%	23800,42	1,9757	61,12%

Tabulka 10: Vyhodnocení pozičního modelu se zahrnutím formy, redukce umístění a tabulkou závislostí

5.5.1 Stíhací závod

Tento druh závodu je oproti všem ostatním specifický, jako jediný totiž nezávisí pouze na výkonu během jediného dne. Závod se totiž startuje Guderssonovou metodou, kde počáteční rozestupy jsou dány výsledky sprintu. Prakticky tedy můžeme předpokládat, že výsledek závisí na závodníkově výkonu v závodu ,ale také výsledku po první části závodu, tedy sprintu.

Aby byla zajištěna obecnost řešení, zavedeme v aplikaci parametr (p) pro poměr důležitosti mezi výsledkem sprintu (startovní pozicí) a zbytkem závodníkových výsledků. Poměr mezi výsledky sprintu a ostatních závodů se definuje vztahem p:1. Celkem závodník absolvoval n závodů a jsou známy i váhy w_i všech jeho výsledků. Váhu w_1 , která představuje význam startovních pozic (tj. výsledky sprintu) určíme ze vzorce (5.10)

$$w_1 = p \cdot \sum_{i=2}^{n} w_i \quad (5.10)$$

Pro otestování výsledného modelu zvolíme p = 1.

Množina dat	ϕ_p	ϕ^2	spi	Max. odchylka
ženy	84,25%	19197,35	2,1056	56,73%
muži	82,69%	27731,65	1,9569	62,81%
Vytrvalostní závod	83,94%	47197,5	2,1213	59,00%
Hromadný start	77,55%	2250,65	1,3662	67,60%
Stíhací závod	87,01%	5111	2,1863	49,01%
Sprint	83,47%	41634,33	2,2128	64,68%
Celkem	83,47%	23464,5	2,0313	59,77%

Tabulka 11: Vyhodnocení pozičního modelu se zahrnutím formy, redukcí umístění, tabulkovou závislostí typů závodů a parametrem pro stíhací závody

6 Model závodních příznaků

Tento model je založen na třech příznacích, které charakterizují každý závod. Jsou jimi rychlost běhu, rychlost střelby a její přesnost. Ke všem příznakům jsou přiděleny váhy stejným způsobem jako v kapitole 5 u pozičního modelu, jedinou změnou je tedy záměna konkrétního příznaku za umístění, které se používalo u předchozího typu modelu. Výsledná křivka, která se opět vypočítá numerickou integrací, je složena z časů jednotlivých příznaků, u všech jsou sledované časy vyjádřeny v základní jednotce [s].

U tohoto modelu již také není potřeba odhadování koeficientů pro stíhací závod, jelikož jsou dosazovány přímo startovní časy.

6.1 Rychlost běhu

Při výpočtech se zanedbává význam závodníkovy techické vyzrálosti, nadmořské výšky i profilu tratí, které jsou popsány v článku 2.3.1. Pro vyhodnocení těchto závislostí není dostatečný počet údajů a výsledky více ovlivňuje aktuální forma závodníků.

Rychlost běhu se získá odečtením celkového času střelby a časem stráveným na trestných okruzích, či penalizacemi. Pro čas strávený na trestném okruhu se pro zjednodušení používá konstanta 25s. Absolutní časy se při každém závodu odlišují v závislosti na délce závodu (i stejné typy závodů se ve vzdálenost mírně odlišují), profilu trati, nadmořské výšce a zejména pak na struktuře sněhu. Není tak možné srovnávat časy mezi jednotlivými závody. Běžeckou výkonnost tak můžeme porovnávat jen v konkrétním závodu. Používané časy jsou tedy definovány jako odstup na nejrychlejšího závodníka. Při relativních odstupech může nastat situace, kde se závodu nezúčastní nejrychlejší závodníci a údaje jsou tak velmi zkreslené. Proto byl zaveden parametr, který pro každý závod stanoví sílu běžecké konkurence a stává se dalším parametrem celkové váhy. Tento parametr je stanoven v každém závodu jako součet umístění závodníků, na startovní listině v klasifikovaném závodu, běžeckých časů patnácti nejlepších v posledních deseti závodech. Výsledná hodnota je následně vydělená číslem 1200, které představuje maximální možný součet (vzorec 6.1) nejlepších běžeckých umístění za předchozích 10 závodů.

$$x = 10 \cdot \sum_{i=1}^{15} i \quad (6.1)$$

Z hodnot rychlosti běhu se vypočte vážený průměr (vzorec 5.7) a vážený rozptyl (vzorec 5.8), z kterých se stanoví pravděpodobnostní rozlišení rychlosti běhu pomocí numerické integrace.

Zkoumaný interval volíme a = 0, b = 1200, počet rozdělení d = 10. Všechny hodnoty následně dosadíme do vzorce (5.6).

Nelze pracovat s běžeckými časy rozdílných závodů, jelikož jejich délka se také liší. Proto se výsledná hodnota t vypočítá pomocí vzorce (6.2), kde t_i představuje běžecký čas v i-tém závodě, l_i délku i-tého závodu a l délku klasifikovaného závodu.

$$t = \frac{t_i \cdot l}{l_i} \quad (6.2)$$

Pro testování je k výpočtu vah použit koeficient formy $\frac{1}{\sqrt{(d_f)}}$, parametr redukce $\sqrt{(d_i)}$ a tabulka závislostí je definována Tabulkou 12. Závod s hromadným startem a stíhací závod nejvíce ovlivňuje v běžecké části taktizování biatlonistů. Proto tyto závody mají pro vytrvalostní závod a sprint nízkou váhu.

	Vytrvalostní závod	Sprint	Stíhací závod	Hromadný závod
Vytrvalostní závod	1	0.75	0.5	0.5
Sprint	0.75	1	0.5	0.25
Stíhací závod	0.75	1	1	1
Hromadný závod	1	1	1	1

Tabulka 12: Tabulka závislostí běžeckých časů

6.2 Rychlost střelby

Rychlost střelby je nutné vyhodnotit pro položku v leže i ve stoje samostnatně. Volbou intervalu a = 15, b = 60 a d = 10 vypočteme pro obě položky z minulých hodnot opět pomocí numerické integrace (5.6) , do které se dosazuje vážený průměr (5.7) a vážený rozptyl (5.8).

K testování dat je váha stanovena pomocí parametru formy $730-(d_f)$, parametr redukce $\sqrt{(d_i)}$ a tabulka závislostí je definována Tabulkou 13. Jelikož přesnost střelby pravděpodobně závisí na konkrétním typu závodu jen minimálně, jsou závislosti u každého závodu stejné. Pouze typ závodu, který odpovídá i klasifikovanému má vyšší váhu. Forma střelby je mnohem stabilnější, než běh, čemuž odpovídá i testovaný parametr.

	Vytrvalostní závod	Sprint	Stíhací závod	Hromadný závod
Vytrvalostní závod	1	0,75	0,75	0,75
Sprint	0,75	1	0,75	0,75
Stíhací závod	0,75	0,75	1	0,75
Hromadný závod	0,75	0,75	0,75	1

Tabulka 13: Tabulka závislostí střeleckých časů

6.3 Přesnost střelby

Také u tohoto parametru určujeme přesnost zvlášť pro oba typy střeleckých položek. Pro střelbu ve stoje i v leže máme dán počet odstřílených položek n, stanovené váhy w_i pro i-tou položku a také počet střeleckých chyb pro i-toupoložku e_i . Úspěšnost střelby se určí dle vzorce (6.3).

$$E(x)=\frac{\sum \frac{e_i}{5}\cdot w_i}{\sum_{i=1}^{n} w_i} \quad (6.3)$$

K testování byly použity stejné podmínky jako pro rychlost střelby.

6.4 Celkový čas

Nejdříve se vypočte průměrný čas, který je dán vzorcem (6.4), kde n představuje počet položek v leže nebo ve stoje (jde o stejné číslo), t_s čas střelby ve stoje, t_l čas střelby v leže, e_s úspěšnost střelby ve stoje, e_l úspěšnost střelby v leže a e_t určuje trest v případě netrefení terče (60s pro individuální závod, pro ostatní 25s).

$$t_a=n\cdot(t_s+t_l+5\cdot((1-e_s)+(1-e_l))\cdot e_t)$$

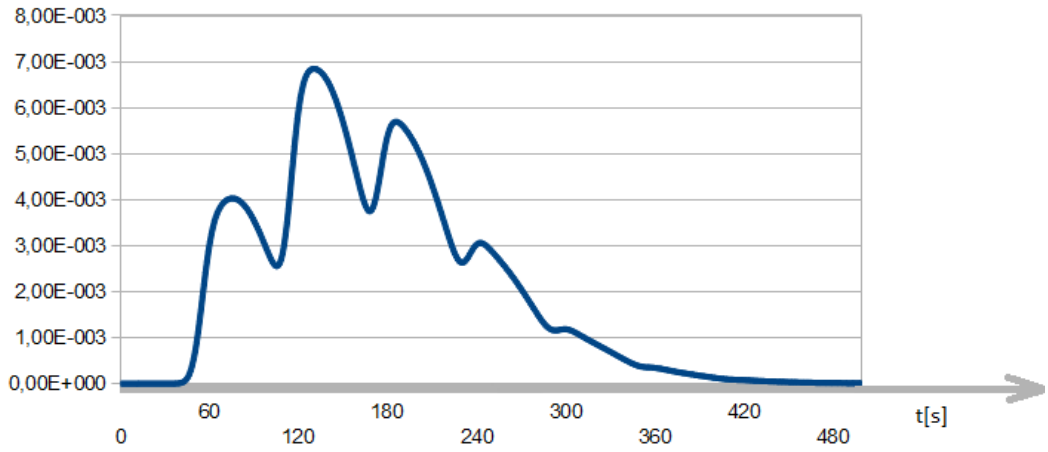
Průměrný čas již řadíme stejně jako v Pozičním modelu umístění, čímž získáme předpokládané umístění závodníků.

Pro určení výsledného časového rozložení každého závodníka je nutné vhodně spojit jednotlivá Gaussova rozložení vyjádřená pomocí numerické integrace přesně podle požadavků klasifikovaného závodu. Pokud jsou obě rozložení v polích a jejich pravděpodobnosti byli děleny po stejných intervalech, vypočte se nové rozlišení pomocí následujícího algoritmu:

```
public static double[] countedProbabilities(double[] probability1, double[] probability2){

    double[] out_probabilities = new double[probability1.length+probability2.length-1];
    for (int i = 0; i < probability1.length; i++){
        for (int j = 0; j < probability2.length; j++){
            if (probability1[i]==0){
                break;
            }
            out_probabilities[i+j] += probability1[i]*probability2[j];
        }
    }
    return out_probabilities;
}
```

Výsledné časové pravděpodobnostní rozložení je získáno postupným slučováním rychlosti běhu, rychlosti střelby a přesnosti střelby přesně dle pravidel pro jednotlivé závody. Příklad rozložení vypadá následujícím způsobem (Obrázek 2).



Výsledná úspěšnost tohoto modelu je znázorněna v Tabulce 14.

Množina dat	ϕ_p	ϕ^2	spi	Max. odchylka
ženy	84,38%	18774,58	2,3212	57,51%
muži	82,85%	27049,46	2,1828	63,36%
Vytrvalostní závod	83,98%	48004,58	2,2154	61,81%
Hromadný start	77,56%	2185,85	1,4642	64,11%
Stíhací závod	87,05%	5283,19	2,6170	52,77%
Sprint	83,80%	39850,4	2,3648	64,31%
Celkem	83,62%	22912,02	2,2520	60,43%

Tabulka 14: Úspěšnost klasifikace

7 Vyhodnocení statistických modelů

7.1 Shrnutí modelů

Nejjednodušším model založený na aritmetickém průměru jak ukazuje Tabulka 15 poskytuje nejméně přesné výsledky ze všech testovaných modelů. Dynamické rozšíření poskytuje zlepšení všech zkoumaných hodnot, v případě volby parametru formy = $1/\sqrt{(d_f)}$ se klasifikace podstatně zlepší, průměrná relativní odchylka o 1,15% oproti modelu s aritmetickým průměrem a o 0,83% oproti modelu s parametrem formy = $(730-d_f)$. Jednoznačně tedy správné zvolení výpočtu formy hraje velkou roli a značně zlepší výslednou úspěšnost.

Obohacení modelu o redukci velkých výkyvů pomocí přidělení nižší váhy horším výsledkům se odchylka od skutečných výsledků prakticky nezlepší, dojde jen k minimálnímu zlepšení, stejně tak u čtercové odchylky umístění a maximální odchylky. Znatelné zlepšení ovšem zaznamenává pravděpodobnostní rozlišení. Tyto výsledky jednoznačně dokazují, že průměrné umístění závodníků zůstává obdobné, model však získává přesnější informace pomocí redukce umístění, což vede ke snížení rozptylů.

Model	ϕ_p	ϕ^2	spi	Max. odchylka
Poziční – aritmetický průměr	81,80%	25169	1,8125	63,07%
Poziční – dyn. rozšíření $(730-d_f)$	82,12%	24542	1,8518	62,50%
Poziční – dyn. rozšíření $\frac{1}{\sqrt{(d_f)}}$	82,95%	23921	1,8778	61,33%
Poziční – obohacení o redukci	83,00%	23847	1,9745	61,16%
Poziční – obohacení závislostí na typu závodů	83,06%	23800	1,9757	61,12%
Poziční – obohacení o koeficient stíhacího závodu	83,47%	23465	2,0313	59,77%
Model závodních příznaků	83,62%	22912	2,2520	60,43%

Tabulka 15: Souhrn úspěšnosti klasifikace modelů

Další model přidává závislost na určitém typu závodu pomocí tabulky závislostí. Ve srovnání s předcházejícím modelem se dají výsledky označit za identické. Jelikož byla určena tabulka závislostí dle zkušeností autora práce, je nutné ověřit, jestli tyto závislosti nebyly určeny špatně a výsledek tak negativně neovlivňuje jen některý z typů závodů, u kterého byly tyto parametry špatně stanoveny. Porovnají se tedy tabulky 8 a 10. Jelikož jsou tabulky velmi podobné a ve většině

sledovaných výsledků došlo k mírnému zlepšení, parametry byly pravděpodobně určeny správně. Rozdíly mezi typy závodů nejsou tedy tak velké, aby jednoznačně měly vliv na umístění závodníků.

Poslední typ pozičního modelu pak řeší problém stíhacího závodu, kde závodníci startují Gundersenovou metodou dle výsledků předchozího závodu (sprintu). Výsledky klasifikace jsou uvedeny v tabulce 11, která dokazuje význam této modifikace a výrazně vylepšuje vypočtené hodnoty.

Model založený na závodních příznacích, (běh, střelba) s výsledky uvedenými v Tabulce 14, mírně vylepšuje průměrnou relativní odchylku i odchylku čtercovou. Maximální odchylka však vzrostla o 0,67%. Úspěšnost pravděpodobnostního rozložení se výrazně zvedá. Zejména u stíhacího závodu, což je dáno přesnými časovými odstupy na startu této disciplíny, s kterými se pak dále počítá.

7.2 Určení nejlepšího modelu

Nejlepší výsledky s výjimkou maximální odchylky přináší model založený na závodních příznacích. Maximální odchylka je však nejméně důležitým parametrem pro vyhodnocení úspěšnosti. Model závodních příznaků však přináší velký pokrok v úspěšnosti pravděpodobnostního rozložení. Právě pravděpodobnostní rozložení je nejdůležitější při stanovení kurzů a tak tento model bezpochyby předčil ostatní.

7.3 Rozdíly úspěšností

Při detailním prohlédnutí všech tabulek informujících o úspěšnosti predikce výsledků lze vypozorovat shodný jev u všech tabulek. U všech zpracovaných hodnot je vyšší úspěšnost klasifikace u závodů žen. Což lze snadno vysvětlit větší vyrovnaností startovního pole mužů, kde se tak obtížněji stanovují přesné výsledky. I podmnožiny dat v případě kategorií vypovídají o stejném pravidlu, nejnižší úspěšnosti jsou jednoznačně zaznamenány u závodu s hromadným startem, kterého se účastní pouze 30 závodníků, kteří jsou vybráni jako aktuálně nejlepší, jsou mezi nimi tudíž i malé rozdíly. Nejvyšší úspěšnost sice dosahuje závod s hromadným startem, kterého se také může účastnit omezený počet závodníků, stanovený na 60. Závod však je odstartován s rozestupy, které se dají klasifikovat a omezují tak i možný náhodný rozptyl závodníků, zapříčiněný nejčastěji chybami na střelnici.

8 Prezentace výsledků a statistik

Požadavky uživatelů neustále rostou. V dnešní době se připojení k internetu pokládá za samozřejmost a spousty aplikací se tak přesouvají na web, uživatel pak může program používat bez ohledu, jestli pracuje na svém osobním počítači, telefonu či jiném zařízení.

8.1 Návrh webové aplikace

Pro webovou aplikaci byl vybrán programovací jazyk PHP, data jsou uložena ve sdílené databázi MySql, která je také využívána hlavní aplikací. O vzhled a uživatelskou funkcionalitu se stará javascript a css.

Webové aplikace jsou si často v mnohém velmi podobné, proto vznikají tzv. Frameworky, které tyto nejčastěji řešené problémy řeší a mají kontrolu nad chodem celého programu. Využit je vlastní framework, který oproti stávajícím řešením vyniká svojí jednoduchostí a rychlostí. Při jeho vývoji byl využit návrhový vzor MVC [5], který odděluje logiku aplikace od její prezentační vrstvy.

Rychlost je zaručena nejen malým počtem funkcí, kvůli kterým obvykle jsou malé frameworky výkonnější, ale především kompilátorem šablon. K sestavení šablon dochází pouze v tzv. vývojovém režimu. Tento framework přidává také kompilátor zdrojového kódu, který pak není sestavován pokaždé při obsluze požadavků a zároveň vkládá pouze knihovní funkce, které jsou v aplikaci volány, ostatní nejsou přiloženy, což snižuje paměťové nároky a umožňuje postavení robustního frameworku bez snižování jeho rychlosti.

8.2 Uživatelé

V rámci aplikace je nutné uživatele rozdělit do několika skupin (administrátor, registrovaný uživatel, neregistrovaný uživatel – host) K tomuto rozdělení je vytvořena funkcionalita [6], která uživateli přiřadí určité oprávnění(role). Každá operace, která vyžaduje oprávnění využívá tzv. Zdroje, k nimž jednotlivé role mají přiřazené operace, které mohou uživatelé vykonávat, na jejich základě se rozhoduje, jestli požadovaná operac bude vykonána.

8.2.1 Registrace

Nejdříve je nutné uživatele vytvořit (zaregistrovat jej), k čemuž je nutné dodržet základní bezpečnostní standardy pro ověřování totožnosti pomocí emailu [7].

Registraci často webové aplikace nahrazují přihlášením pomocí sociální sítě facebook [8].

Tato přihlašování může být pro uživatele pohodlné, nicméně pro web, kde potřebujeme od uživatele získat citlivé informace a připravit web i na implementaci platebních metod, je tento přístup nevhodný.

8.3 Bezpečnost aplikace

Pro každou aplikaci, zejména pak webové, hraje možná nejpodstatnější roli její bezpečnost a také dostupnost. Pokud služba nabízí zajímavé služby ale mívá časté výpadky, uživatelé si rychle naleznou jejich alternativu i za cenu nižší přidané hodnoty. Nejhorším možným ale stále ještě velmi častým problémem webových aplikací je jejich bezpečnost, proto je nutné celou aplikaci zabezpečit, zejména proti známým útokům [9].

8.4 Statistiky a predikce výsledků

Hlavní význam této webové aplikace tkví v zobrazování statistik nasbíraných dat a rovněž zobrazení předpokládaných výsledků, které spočítá samotný model a předá webové aplikaci prostřednictvím sdílené databáze.

Přehlednost a uživatelskou přívětivost statistikám dodávají grafy, které jsou generovány pomocí javascriptové knihovny Google Charts [10].

9 Shrnutí a další perspektivy práce

Cílem této práce bylo vyvinout nástroj pro predikci biatlonových výsledků, navrhnout základní statistické modely a výsledky spolu se statistikami prezentovat ve formě webové aplikace.

Nejprve byl naprogramován parser pro stáhnutí dat z oficiálního serveru biatlhlonworld.com. Převádí soubory z textové podoby ve formátu pdf do datavé struktury a následně je ukládá do databáze MySql. K vyhodnocování výsledků se použilo mimo standartně používaných odchylek i pravděpodobnostní rozložení, které se skládá zpravidla z několika Gaussových křivek, vyjádřených pomocí numerické integrace. Dále byl vytvořen základní model, nazývaný jako Poziční, který závisí pouze na umístění závodníků v přecházejících závodech. Nejjednodušší způsob vyhodnocení výsledků tímto modelem je určením aritmetického průměru, tento postup však nepřináší dostatečně přesné výsledky a tak byl model dynamicky rozšířen. Toto rozšíření vyjadřuje závislost výsledků závodníků, na jejich aktuální výkonnosti, která je převedena na váhu výsledků a určena pomocí data konání závodů. Dynamické rozšíření se ukázalo jako velmi přínosné a dokázalo tak, že výkony závodníků se během sezóny postupně výrazně mění. Váha nakonec byla obohacena ještě o parametr redukce nejhorších (nejlepších) výsledků a tabulkovým vyjádřením závislostí typů závodů. V konkrétních modelech však tyto modifikace nepřinášely výrazné zlepšení výsledků, další zřejmý posun tak znamenalo až zavedení koeficientu pro stíhací závod, výsledky tohoto typu závodu se zlepšili ve většině parametrů až o 2 procenta.

Další typ modelů byl vytvořen na základě závodních příznaků (běh, střelba). Tento model díky obecnému návrhu vah mohl přejmout zapsané funkce od pozičního modelu se záměnou za nové příznaky místo umístění. Běžecké časy není možné porovnávat mezi různými závody a určit tak nejlepší běžce. Všechny časy se počítají relativně k vítěznému času, což znemná stanovit i kvalitu vítězného času pomocí určení síly konkurence ve vybraném závodu. Model založený na závodních příznacích prokazuje jednoznačně nejlepší úspěšnost klasifikace, zejména pak pravděpodobnostní rozložení, potřebné ke stanovení kurzů.

Všechny statistiky, které parser získá jsou k dispozici přes sdílenou databázi webové aplikace, která umožňuje zobrazení vybraných závislostí a usnadňuje bookmakerovi či sázkaři vytvořit si přehled o závodníkovi. Aplikace oproti oficiálním statistikám generuje i grafy běžeckých časů. Všechny statistiky umožňuje porovnávat v jednom grafu přidáním dalších závodníků.

Na tuto práci je možné dále navázat a pomocí matematických výpočtů přesněji určit parametry, které byly v některých případech pouze určeny ze zkušeností autora této práce. Díky zavedení modelu se závodními příznaky bude mírnou úpravou možné velmi snadno vyhodnocovat

závod i po každém odběhnutém okruhu. Dále je potřeba přidat výpočet, který při malém počtu dostupných informací přiblíží jednotlivé závodní příznaky blíže průměrným, snadno by se totiž mohlo stát, že závodník s jedním absolvovaným závodem získá 100% úspěšnost střelby. Současné kompletní vyhodnocení výsledků u druhého z modelů na běžném počítači snadno přesáhne 10 minut, proto by bylo vhodné pro výpočty ještě více využít ukládání parametrů, které se často opětovně vyhodnocují. Program by při nižší náročnosti mohl být užit i jako samoučící a sám by se snažil pomocí náhodného (nebo omezeně náhodného) generování parametrů přijít na kombinace zajišťující lepší výsledky klasifikace.

Seznam použité literatury

[1] Der Glücksspielmarkt. BET-AT-HOME.COM. *Investor relations* [online]. [cit. 2013-04-08]. Dostupné z: http://www.bet-at-home.ag/Default.aspx?page=4

[2] Bwin Sports Bets. BWIN GROUP. *Bwin pressevent* [online]. [cit. 2013-04-15]. Dostupné z: http://www.bwinpressevent.com/images/bwin/bwin/sportsbets_en.pdf

[3] HEJDUŠEK, Marek. *Matematické metody pro modelování a predikci výsledků sportovnıch utkání*. Praha, 26.11.2006 [cit. 2013-05-19]. Diplomová práce. CVUT. Vedoucí práce doc. Petr Volf, CSC.

[4] IBU EVENT AND COMPETITION RULES. [online]. 1998, 2012 [cit. 2013-05-12]. Dostupné z: http://www.biathlonworld.com/media/files/downloads/IBU_Rules_2012_cap3.pdf

[5] Model View Controller Architercture. [online]. 6.7.2010. [cit. 2013-05-10]. Dostupné z: http://www.codeproject.com/Articles/92182/Model-View-Controller-Architecture

[6] HLINA, Igor. Statické ACL v modulárnej aplikácii. GRUDL, David. *Nette Framework* [online]. 1. 2. 2013 [cit. 2013-05-15]. Dostupné z: http://pla.nette.org/cs/staticke-acl

[7] Kontrola e-mailové adresy. VRÁNA, Jakub. *PHP triky* [online]. 5.5.2006. [cit. 2013-05-19]. Dostupné z: http://php.vrana.cz/kontrola-e-mailove-adresy.php

[8] Přihlášení přes Facebook. *Tvorba webů, programování a business dohromady* [online]. 7.4.2012 [cit. 2013-05-16]. Dostupné z: http://www.webusiness.cz/2012/04/prihlaseni-pres-facebook/

[9] FRESCHMANN, Petr. Bezpečnost na webu – přehled útoků na webové aplikace. *Zdroják* [online]. 10.11.2008 [cit. 2013-05-16]. Dostupné z: http://www.zdrojak.cz/clanky/prehled-utoku-na-webove-aplikace/

[10] Google Charts. GOOGLE. *Google* [online]. 3.4.2012 [cit. 2013-05-16]. Dostupné z: https://developers.google.com/chart/

Příloha A

Obsah CD

- Bakalářská práce ve formátu PDF
- Struktura databáze, data za poslední 3 sezóny
- Zdrojový kód webové aplikace (PHP) vč. dokumentace
- Zdrojový kód statistických funkcí (Java) vč. dokumentace
- Manuál ke zprovoznění aplikací